ANOMALY DETECTION WITH SELF-SUPERVISED AUDIO EMBEDDINGS

Technical Report

Ivan Zorin

Artificial Intelligence Research Institute (AIRI)
Industrial AI
Moscow, Russia
zorin@airi.net

Ilya Makarov

Artificial Intelligence Research Institute (AIRI)
Industrial AI
Moscow, Russia
makarov@airi.net

ABSTRACT

The majority of approaches to machine condition monitoring via anomalous sound detection are based on supervised learning. The metadata of the datasets is used as data labels for training supervised models. However, data labeling is expensive and often impossible for industries with significant amount of equipment. In this case self-supervised methods could solve the problem since they do not require labeled data. In this work we applied the recent self-supervised approach to compute embeddings of audio signals named BYOL-A and classical machine learning method Local Outlier Factor (LOF) to compute outlier scores for anomalous sounds. The main focus of this work is to not use any labels from the metadata of the datasets and explore a self-supervised learning approach.

Index Terms— Self-supervised learning, anomaly detection

1. INTRODUCTION

The goal of acoustic anomaly detection is to determine, analyzing an audio recording, whether a machine recorded works in normal regime or it is malfunctioning producing anomalous sound. The problem of industrial machine monitoring has significant importance to the industry since it can prevent malfunctioning of expensive equipment. Early detection of anomalous working regime decreases the time out of service and therefore decreases expenses. There are different types of sensors that can be installed on a machine for monitoring purposes, such as vibration and acoustic microphones. There are two main advantages of using microphones as monitoring sensors, first, they are considerably cheaper than other types of sensors, and secondly, they can be installed noninvasively, which makes them a universal sensor for almost any kind of industrial machines.

Although the Task 2 is formulated as self-supervised learning (SSL) [1], the majority of submissions of the previous editions used supervised approaches. These submissions trained classifiers to predict labels from metadata of the challenge dataset, like machine ID or working mode, and then extracted embeddings from a layer before the classification head. In a certain sense, using metadata can be considered as a data leak. The main problem with this approach is that labeling during data collection is required, which may be impossible to get in a real-world industrial site scale. In real cases, dataset can be represented by a set of unlabeled files (at best recorded from a known machine type) of normal working conditions and a significantly smaller set of anomalous working conditions. Therefore, it can be impossible to train a model in a supervised manner using metadata.

Summing up, there is a need of a good self-supervised solution. We focus our attention on models for self-supervised representation of speech signals. There are several successful applications of self-supervised speech representation for Automatic Speech Recognition (ASR) task, wav2vec [2] and wav2vec 2.0 [3], HuBERT [4], TERA [5], BYOL-A [6] to name a few. The main idea is twofold. First, we use classic self-supervised training with Contrastive Predictive Coding loss or its variants to learn hidden representation of the speech data. Second, we use these representations as the input features for the downstream task instead of classical Short-time Fourier transform (STFT) or log-Mel filterbanks features.

2. PROBLEM STATEMENT

We formulate the problem as a self-supervised learning task. Our model was trained on data of normal working regimes, which is motivated by two factors. First, there is a significant disproportion of normal and anomalous training data since anomalies are rare events. Second, the choice is motivated by the industrial needs in a solution that can be adapted to different types of machines and their working regimes. On a nutshell, our approach is based on

(i) training an embedding model in a self-supervised approach;

(ii) embedding normal and anomalous audios;

(iii) fitting a Local Outlier Factor model to score embeddings and predict outliers, i.e. anomalous conditions.

2.1. Embedding Model

To compute embeddings, we looked at two SSL approaches. The first one is based on Contrastive Predictive Coding, in which a model is trained to learn a context of an input sequence and predict next sample of it using the learned context. Then the context is used as the embedding. The second one uses the idea that embeddings of the same file with different augmentations should be close in the embedding space. We picked the second one following the intuition that the first one is less suitable for industrial audio domain since it founds its success in a speech domain where the context is to vary more comparing to industrial sounds.

We used a model called BYOL-A, which is a modification of computer vision model BYOL [7] for audio domain, to compute embeddings, with architecture as it was reported in the paper and

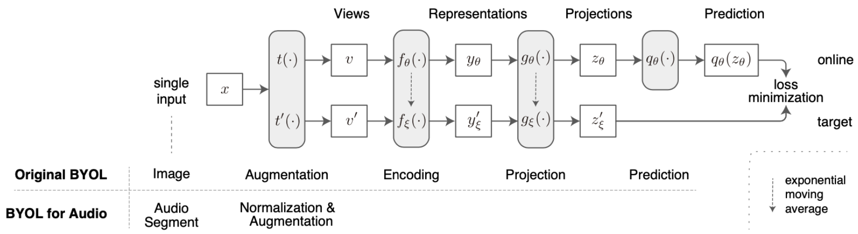

Figure 1: Pipeline of BYOL model

the code open sourced by the authors. The input features of the model are log-Mel spectrogram computed from the whole length audio files with 1024 points of Fourier Transform and 64 Mel filters. Log-mel spectrogram then goes through the augmentation block of mixup and random crop and resize, with pre- and post- mean-variance normalization. We used only original datasets to train the model [8] and [9]. It is important to highlight that we did not use any labels of the files from the datasets, because it can be considered as a data leak, and moreover such labels are not available in real-world cases. Figure 1 overviews the model.

2.2. Anomaly scoring

To compute anomaly scores for each input file, we fitted Local Outlier Factor on the embeddings of the normal files from train set. Then embeddings for the test set files were computed and scored with the Local Outlier Factor instance.

2.3. Quality Metrics

The main quality metric used in this study is Area Under the receiver operating characteristic Curve (AUC). The formulas to compute AUC and pAUC are Equation 1 and Equation 2, respectively.

$$AUC = \frac{1}{N_- N_+} \sum_{i-1}^{N_-} \sum_{j=1}^{N_+} \mathbb{I}[\mathcal{A}_\theta(x_j^+) > \mathcal{A}_\theta(x_i^-)] \qquad (1)$$

$$pAUC = \frac{1}{\lfloor pN_- \rfloor N_+} \sum_{i-1}^{\lfloor pN_- \rfloor} \sum_{j=1}^{N_+} \mathbb{I}[\mathcal{A}_\theta(x_j^+) > \mathcal{A}_\theta(x_i^-)] \qquad (2)$$

where $\{x_j^+\}_{j=1}^{N_+}$ and $\{x_i^-\}_{i=1}^{N_-}$ are anomalous and normal sets of signals, respectively; $\mathcal{A}_\theta(x)$ is a scoring function. We took $p = 0.1$ in our experiments.

3. RESULTS

The final results are presented in Table 1. Here we do not separate machines by their sections since we followed fully self-supervised logic and did not use any metadata while training. The results in Table 2 were computed for each machine type over all its files together, without splitting by the section IDs. Table 2 presents results for machines separately per section IDs, as it was reported on the challenge website. In our view, Table 2 is more representative for self-supervised problem formulation, because it is consistent with the idea to not use the labels of the audio files.

Figure 2 shows the embeddings of two machines, gearbox and slider, from the dataset after TSNE dimensionality reduction. As one can see on the plot, certain sections of training data were densely clustered.

Table 1: Resulting AUC on *dev_test* dataset

machine	gearbox	valve	slider	fan	ToyTrain	ToyCar
AUC	0.57	0.50	0.65	0.54	0.60	0.54

Table 2: AUC and pAUC results by machine type and section ID

section	**source domain AUC**	**target domain AUC**	**pAUC**
gearbox			
section 00	0.56	0.54	0.49
section 01	0.65	0.496	0.51
section 02	0.64	0.50	0.50
harmonic mean	0.62	0.51	0.50
valve			
section 00	0.60	0.64	0.55
section 01	0.49	0.50	0.49
section 02	0.50	0.30	0.50
harmonic mean	0.52	0.43	0.51
slider			
section 00	0.75	0.62	0.54
section 01	0.62	0.51	0.53
section 02	0.71	0.63	0.53
harmonic mean	0.69	0.58	0.53
fan			
section 00	0.56	0.48	0.49
section 01	0.52	0.59	0.53
section 02	0.51	0.62	0.49
harmonic mean	0.53	0.56	0.50
bearing			
section 00	0.41	0.54	0.48
section 01	0.40	0.46	0.49
section 02	0.54	0.47	0.48
harmonic mean	0.44	0.49	0.49
ToyTrain			
section 00	0.56	0.58	0.51
section 01	0.57	0.63	0.52
section 02	0.73	0.58	0.51
harmonic mean	0.61	0.60	0.51
ToyCar			
section 00	0.63	0.53	0.52
section 01	0.53	0.40	0.53
section 02	0.67	0.55	0.54
harmonic mean	0.61	0.49	0.53

4. CONCLUSION

In this submission, we tested the possibility of detecting anomalous machine conditions using fully self-supervised approach building embeddings of audio files and then detecting outliers of these embeddings distribution. The model was trained with files of normal working conditions only without any metadata labels, such as section ID or operational settings. We believe that our results can be a baseline for fully self-supervised approaches in further researches. We point two possible directions for further research, the first one is to experiment with the embedding model and outlier detector to improve the results. The second is to train a universal model invariant to a machine type, which can make an industry-ready solution.

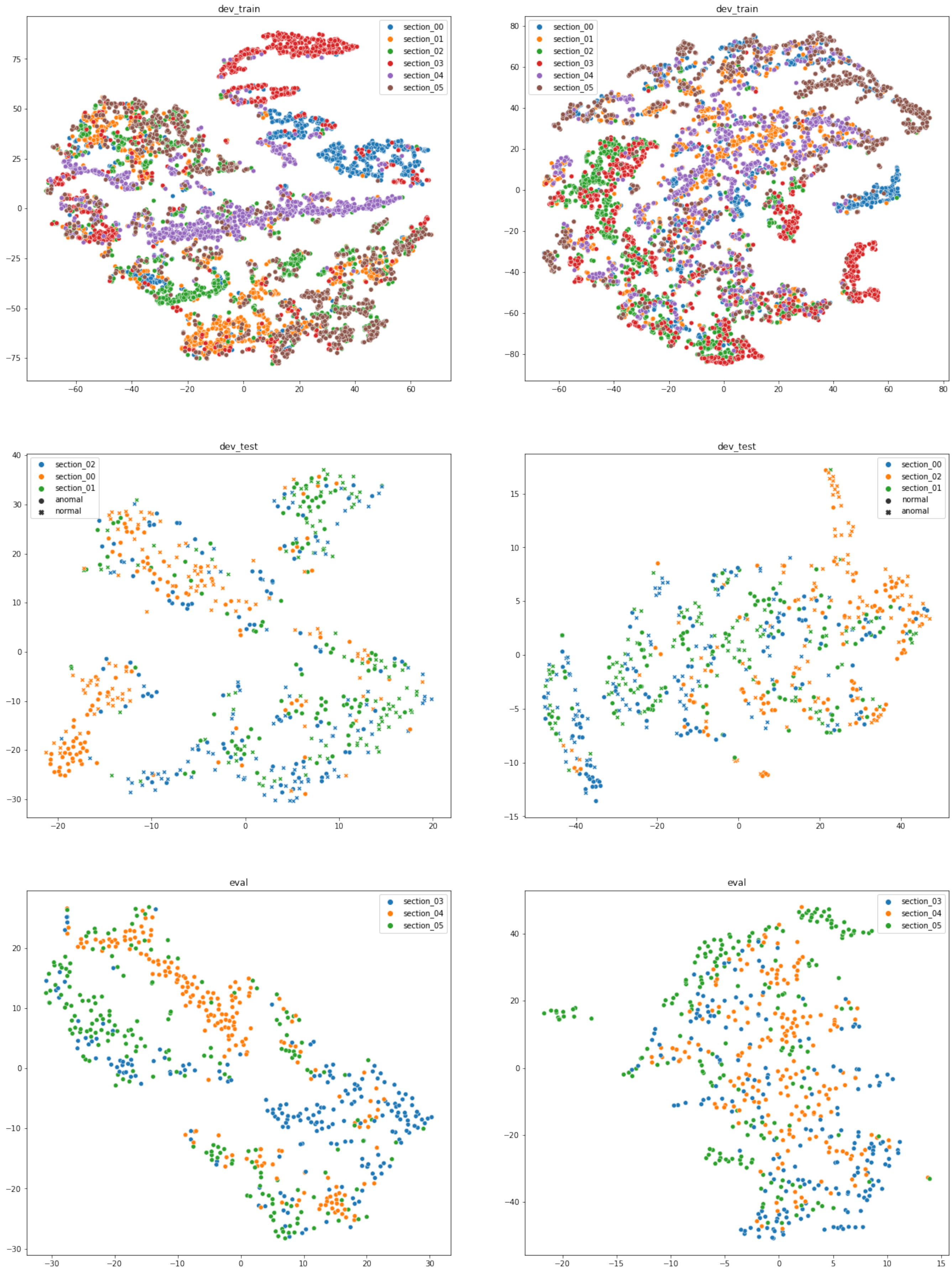

Figure 2: TSNE visualization of embeddings of *dev_test*, *dev_train* and *eval* datasets of gearbox (left) and slider (right)

5. REFERENCES

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *In arXiv e-prints: 2206.05876*, 2022.

[2] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[6] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[7] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[8] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," *In arXiv e-prints: 2205.13879*, 2022.

[9] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 1–5.